BCORLE(λ): An Offline Reinforcement Learning and Evaluation Framework for Coupons Allocation in E-commerce Market

Yang Zhang，Bo Tang, Qingyu Yang，Dou An，
Hongyin Tang, Chenyang Xi, Xueying Li, Feiyu Xiong
Xi’an Jiaotong University & Alibaba Group

Title

CONTENT

01 Background

PART ONE

Coupons allocation

(a) The App screen view of the main page of Taobao Deals

(b) The App screen view of the daily check-in scenario of Taobao Deals

Requirements:

1. Maximizing users' retention
2. Prevent the cost from exceeding the budget
3. Respond quickly to the changing business strategy

Previous work

Uplift model

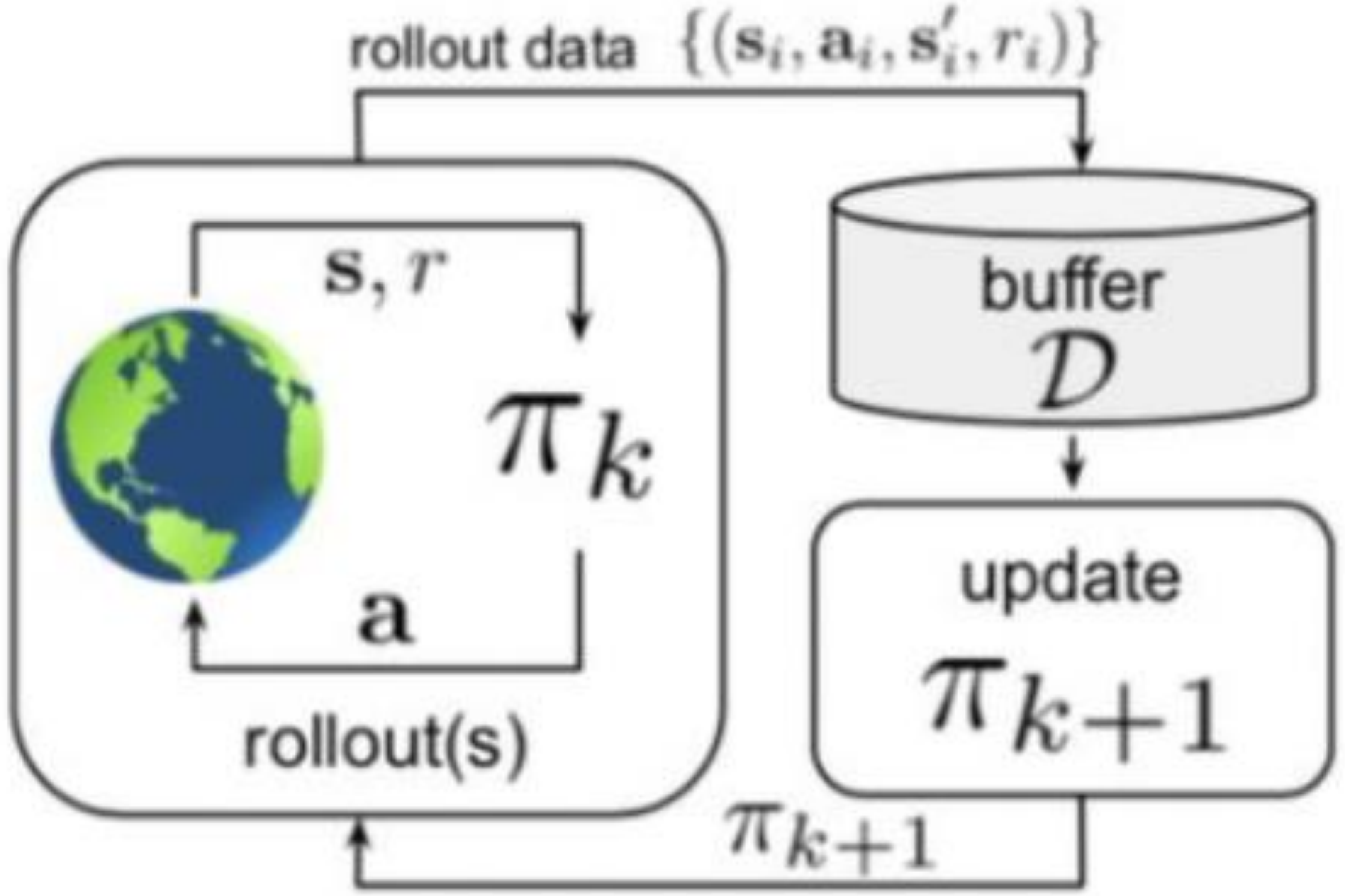
- ✓ Predicting the uses' retention intent after receiving different coupons
  Method: Logistics regression, Gradient boost decision tree
- ✓ Action selection of coupons allocation
  Method: Linear programming

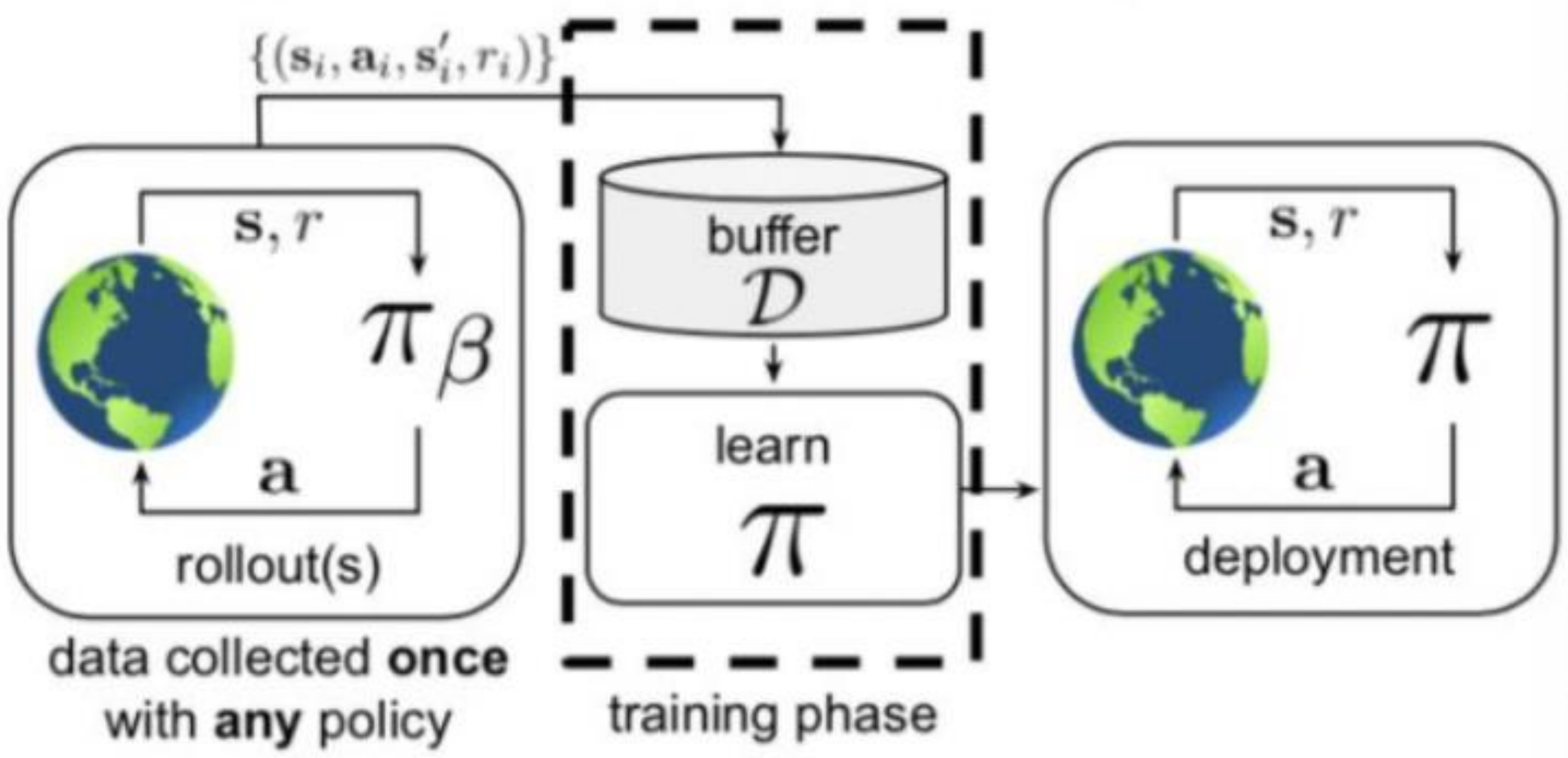
Disadvantages

- ✓ Aim to maximize the benefit in one day
- ✓ Ignore the future benefits in making decision

The graph comes from "Google Research Football: A Novel Reinforcement Learning Environment" [AAAI'2020]

RL → Offline RL





✓ Maximize the benefits during one episode of days

✓ Train the policy in an offline manner

Our work

- Budget constrained offline reinforcement learning and evaluation with λ-generalization (BCORLE(λ))
  - λ-generalization method
  - Resemble batch-constrained Q-learning (R-BCQ)
  - Random ensemble mixture evaluator (REM)

02 Core work

PART TWO

Problem formulation

Objective function

$$\max_{\pi\in\Pi} J(\pi) = \mathbb{E}_{\tau\sim\pi,\mu}\left[\sum_{t=0}^{T}\gamma^t r_t\right]$$

Subject to

$$C(\pi) = \mathbb{E}_{\tau\sim\pi,\mu}\left[\sum_{t=0}^{T}\gamma^t c_t\right] \leq b$$

Convert to Lagrangian problem

Objective function

$$\max_{\pi\in\Pi} L(\pi,\lambda) = J(\pi) - \lambda(C(\pi) - b)$$

Subject to

$$\lambda \geq 0$$

Theorem analysis

Assumption 1. *There exists a policy π that satisfies the constraint* $C(\pi) < b$

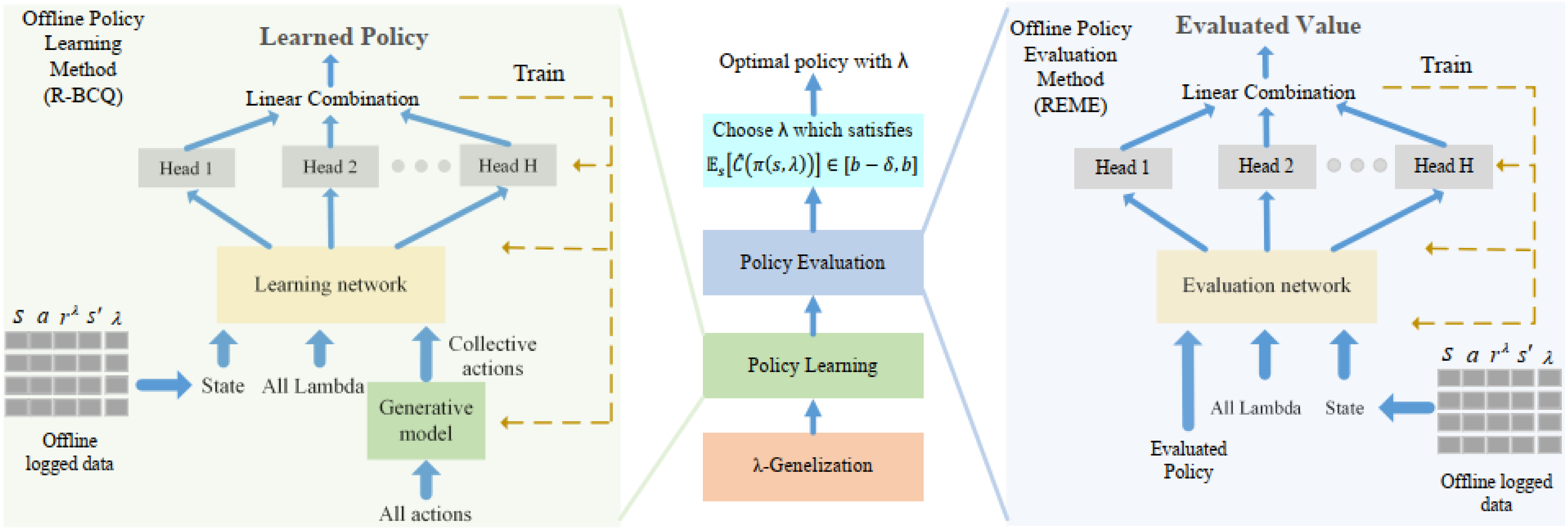
Assumption 2. *Given* λ_a *and* λ_b *, if* $C(\pi^*_{\lambda_a}) \geq C(\pi^*_{\lambda_b})$ *, then* $J(\pi^*_{\lambda_a}) \geq J(\pi^*_{\lambda_b})$*, where* $\pi^*_{\lambda_a} = \arg\max_\pi L(\pi, \lambda_a)$ *and* $\pi^*_{\lambda_b} = \arg\max_\pi L(\pi, \lambda_b)$ *.*

Assumption 3. *There exists* λ_a *and* λ_b*, making the condition* $C(\pi^*_{\lambda_a}) < b$ *and* $C(\pi^*_{\lambda_b}) > b$ *hold.*

Theorem 1. $C(\pi^*_\lambda)$ *is monotonically non-increased with the increase of λ, i.e., If* $\lambda_a \leq \lambda_b$ *, then* $C(\pi^*_{\lambda_a}) \geq C(\pi^*_{\lambda_b})$ *.*

Theorem 2. *Under Assumption 2 and Assumption 3, their exists an optimal Lagrangian multiple variable* λ^* *which can make its corresponding optimal policy* $\pi^*_{\lambda^*}$ *maximize the objective function J(π) while satisfying the budget constraint.*

Our approach: BCORLE(λ) framework



λ-Generalization

Lagrangian problem —transformed→ RL problem

$$L(\pi, \lambda) = J(\pi) - \lambda * (C(\pi) - b))$$

$$= \mathbb{E}\left[\sum_{t=1}^{T} r(s_t, a_t)\right] - \lambda\left(\mathbb{E}\left[\sum_{t=1}^{T} c(s_t, a_t)\right] - b\right)$$

$$= \mathbb{E}\left[\sum_{t=1}^{T} r(s_t, a_t) - \lambda c(s_t, a_t)\right] + \lambda b$$

$$= \mathbb{E}\left[\sum_{t=1}^{T} r^{\lambda}(s_t, a_t)\right] + \lambda b.$$

The reward function to be optimized: $r^{\lambda} = r(s_t, a_t) - \lambda c(s_t, a_t)$

λ-Generalization

1. Enlarge the transition tuples

$$(s_i, a_i, r_i, c_i, s_{i+1}) \Rightarrow \{(s_i, a_i, r_i^{\lambda_j}, c_i, s_{i+1}, \lambda_j)\}_{j=1}^{L}$$

where $\lambda_j \in \{\lambda_1, \lambda_2, \ldots, \lambda_L\}$

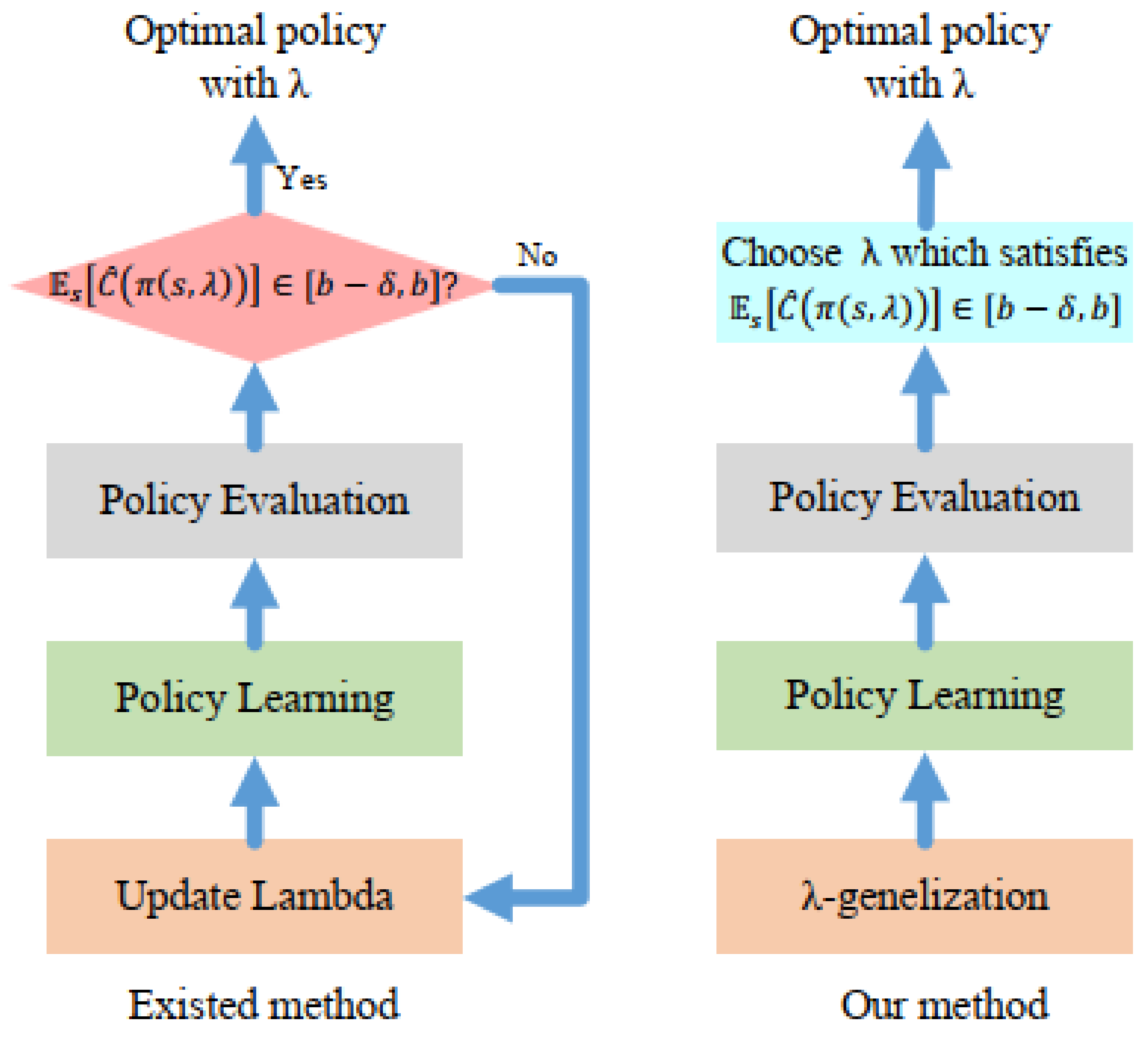
2. Enlarge the training datasets

$$D = \{(s_i, a_i, r_i, c_i, s_{i+1})\}_{i=1}^{M}$$

$$\Downarrow$$

$$D' = \{(s_i, a_i, r_i^{\lambda_j}, c_i, s_{i+1}, \lambda_j)\}_{i,j=1,1}^{M,L}$$

3. Train the policy with different values of λ



XI'AN JIAOTONG UNIVERSITY

Policy training: R-BCQ

- ✓ A generative model $G(a|s,\lambda;\omega)$ to drop state-action pairs which seldom appear in training dataset when selecting action. It is proposed to address the mismatch problem which causes that the estimated value of some state-action pairs deviate greatly from the true value.
- ✓ A multi-head network is employed to increase the robustness and generalization ability of the policy learning network and avoids the Q-value estimation bias problem, compared with common one-head network.
- ✓ Learned policy:

$$\pi(s,\lambda) = \mathop{\arg\max}_{a|G(a|s,\lambda;\omega)/\max_{\hat{a}} G(\hat{a}|s,\lambda;\omega)>\beta} \sum_i \alpha_i Q_i(s,a,\lambda)$$

- ✓ Policy training:

$$L(\theta) = \mathbb{E}_{s,a,r^\lambda,s',\lambda\sim D'}\left[l_\kappa\left(r^\lambda + \gamma \max_{a'|G(a'|s')/\max G(\hat{a}|s')>\beta} \sum_i \alpha_i Q_i'(s',a',\lambda) - \sum_i \alpha_i Q_i(s,a,\lambda)\right)\right]$$

Policy evaluation: REME

- ✓ Employing the *policy π* when calculating the target evaluated value, unlike using the *max operator* in the policy learning.

- ✓ A multi-head network is employed increases the robustness and generalization ability of the policy evaluation network.

- ✓ The update of policy evaluation network:

$$L(\hat{\theta}) = \mathbb{E}_{s,a,r,s',\lambda\sim D'}\left[l_\kappa\left(r+\gamma\sum_i \alpha_i \widehat{Q}_i\left(s',\pi\left(s',\lambda\right),\lambda\right)-\sum_i \alpha_i \widehat{Q}_i\left(s,a,\lambda\right)\right)\right]$$

03 Experiment results

PART Three

Experiment methodology

Studied problem:

- ✓ Does λ-generalization method help to reduce the computation overhead of policy training?
- ✓ How does BCORLE(λ) framework with R-BCQ algorithm perform in comparison to other state-of-the-art offline RL algorithms?
- ✓ How does REME algorithm perform in comparison to other OPE algorithms?
- ✓ How do different values of λ in λ-generalization method affect the performance of proposed approach?

Simulation experiments

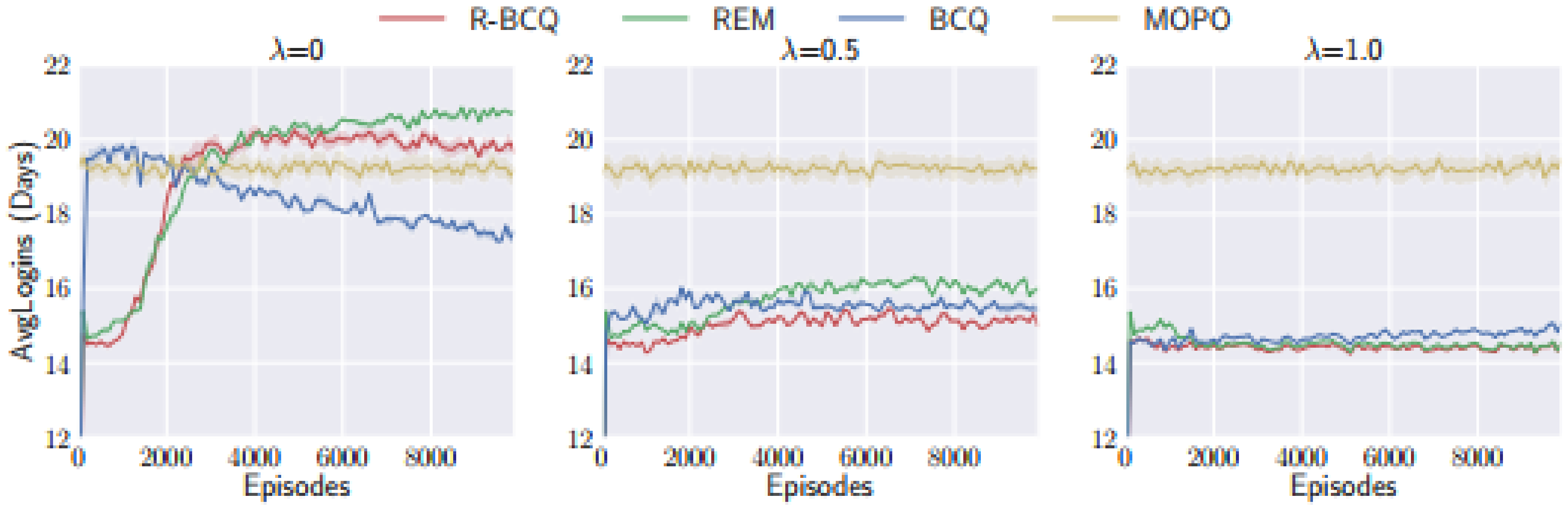
Aim:

ensure there is no risk or unaffordable cost when using the proposed method.

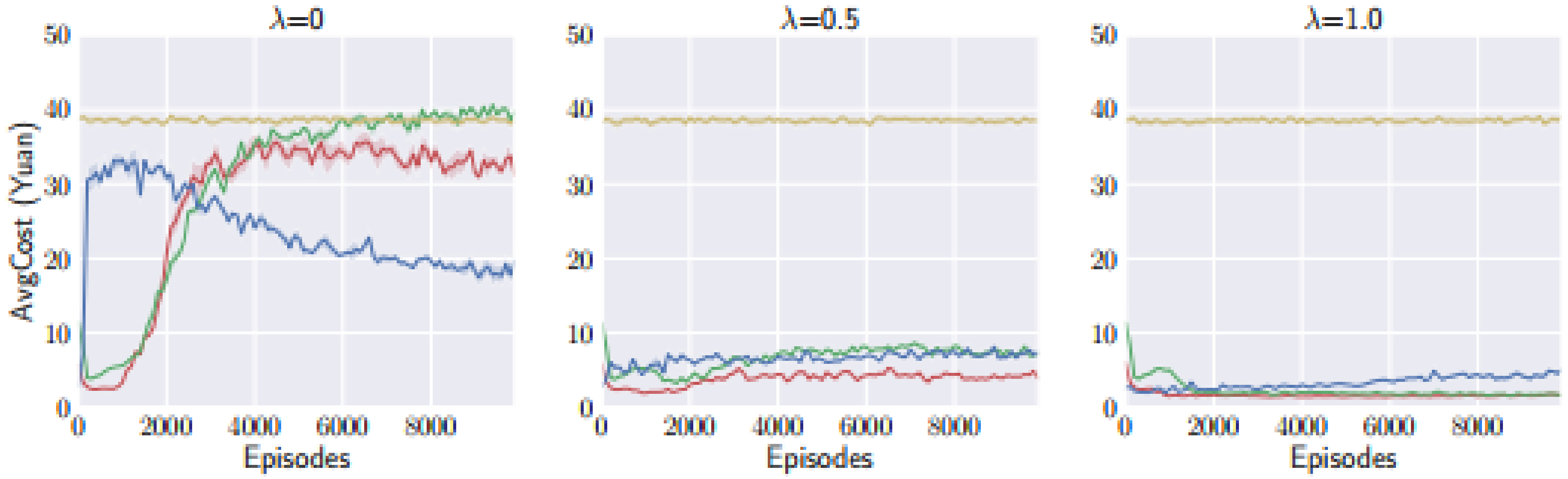
Setup:

- ✓ 10000 users
- ✓ Each user will receive a coupon after logging
- ✓ Each user logs into the simulation platform according to the user preference
- ✓ Time span : 30 days
- ✓ Action type : 21 items of coupons
- ✓ λ values : 0, 0.05, 0.1, … , 0.95, 1.0
- ✓ Reward : 1 when the user logs into the platform, 0 else.

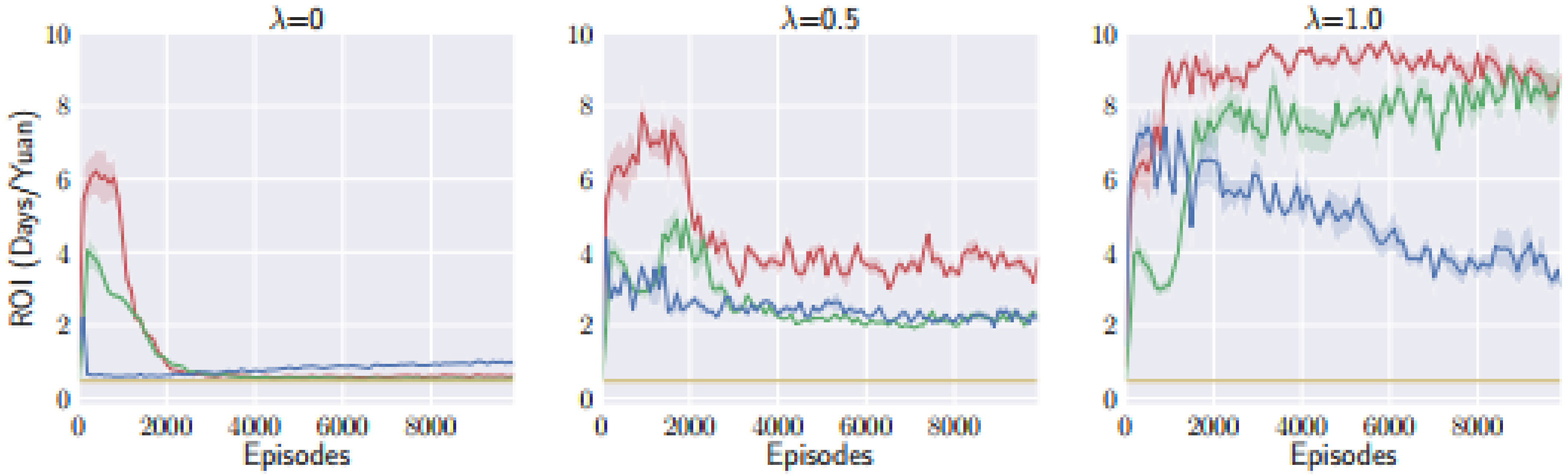
Simulation experiments



(a) The comparison results of AvgLogins



(b) The comparison results of AvgCost



(c) The comparison results of ROI

Method	Episodes	Time cost
λ-update approach	14000	2.187h
Our approach	3000	0.439h

The comparison results of computation overhead

OPE Method	Errors (Days) when $\lambda = 0$			Errors (Days) when $\lambda = 0.5$			Errors (Days) when $\lambda = 1$		
	ls=1000	ls=2000	ls=4000	ls=1000	ls=2000	ls=4000	ls=1000	ls=2000	ls=4000
IS	3.7837	3.1353	2.1883	2.7559	2.9531	3.2068	0.4299	0.4299	0.4301
DM	0.0201	0.0204	0.0191	0.0204	0.0201	0.0195	0.0192	0.0197	0.0194
DR	1.0126	0.8014	0.5139	2.1328	2.2951	2.5462	0.2258	0.2252	0.2258
FQE	0.2064	0.1483	0.0644	0.2102	0.1844	0.1686	0.0196	0.0192	0.0193
REME	**0.0135**	**0.0127**	**0.0158**	**0.0118**	**0.0085**	**0.0093**	**0.0100**	**0.0082**	**0.0091**

The errors of evaluated values

西安交通大學
XI'AN JIAOTONG UNIVERSITY

Real-world experiments

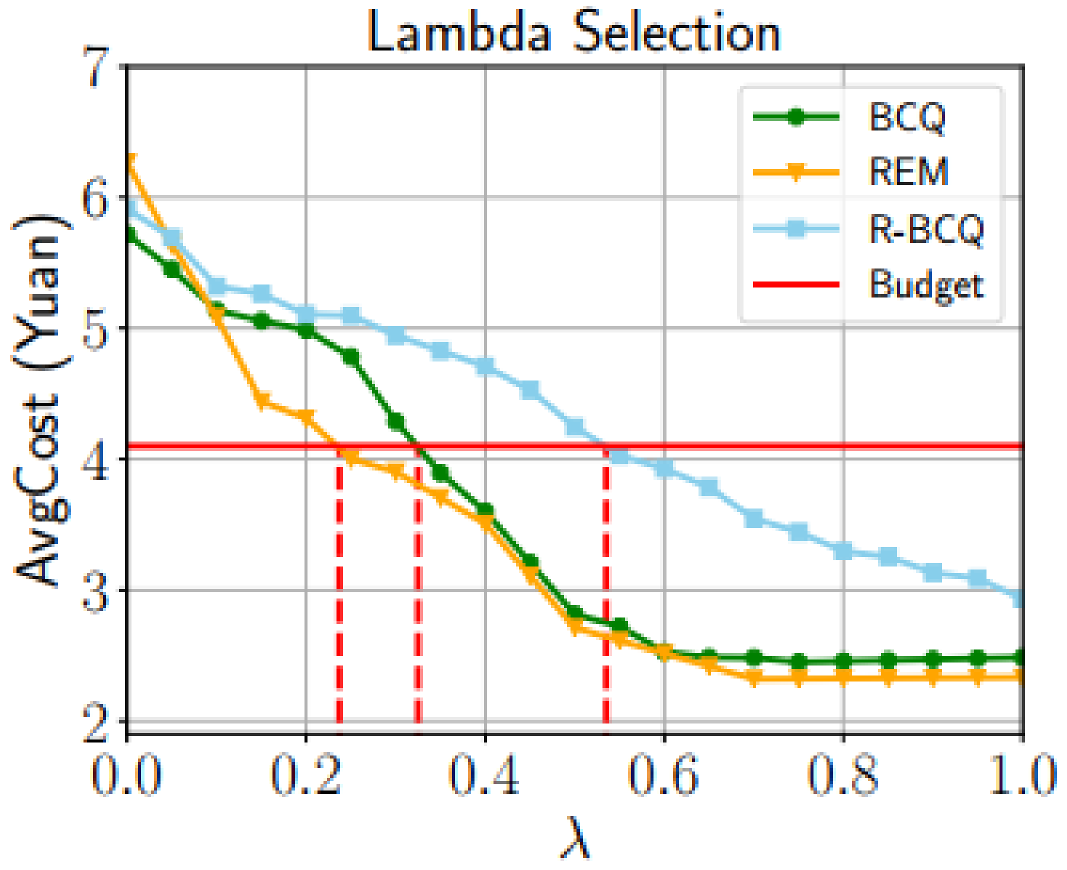
Real-world platform:

Taobao Deals, which is a mobile shopping app launched by Alibaba Group in 2020 with over 10 million daily active users.

Setup:

- ✓ Datasets: over 2 million users' daily check-in records
- ✓ Time span : 14 days from March 9th to March 22nd.
- ✓ Action type : 13 items of coupons
- ✓ The length of one episode : 7 days
- ✓ λ values : 0, 0.05, 0.1, … , 0.95, 1.0
- ✓ Reward : 0 when the user logs into the platform, -1 else
- ✓ Budget : 4.1 Yuan
- ✓ The value of δ : 0.1 Yuan

Real-world experiments



Selecting the policy with λ value

OPE Method	Errors of AvgLogins (Days)	Errors of AvgCost (Yuan)
IS	0.2203	0.2754
DM	0.1417	0.1923
DR	0.1848	0.1881
FQE	0.1933	0.1515
REME	**0.0443**	**0.0221**

The errors of evaluated values

Method	Results in first week				Results in second week			
	AvgLogins (Days)	AvgCost (Yuan)	ROI (Days/Yuan)	ROI Imp	AvgLogins (Days)	AvgCost (Yuan)	ROI (Days/Yuan)	ROI Imp
LR+LP	5.0416	**4.0628**	1.2409	0	5.3099	4.0684	1.3052	0
GBDT+LP	5.0442	4.0776	1.2371	-0.31%	5.3802	4.0756	1.3201	1.14%
BCQ	5.6832	4.0740	1.3950	12.42%	5.8789	4.0698	1.4445	10.67%
REM	5.7108	4.0644	1.4051	13.23%	5.9092	4.0729	1.4509	11.16%
R-BCQ	**5.8252**	4.0654	**1.4329**	**15.47%**	**5.9871**	**4.0528**	**1.4773**	**13.18%**

The online results in Taobao Special Offer Edition app during two weeks.

Thanks

zy804703098@stu. xjtu. edu. cn

Recruitment Information

Welcome to Join Us

—— Algorithm Team in Alibaba Group

Contact： xiaoming.lxy@alibaba-inc.com

美团外卖广告组2022校园招聘——强化学习/运筹优化方向

- 团队介绍

外卖技术部承担外卖业务的技术开发工作，包括用户洞察、智能交互及交易引导、商家经营、平台运营等方面的系统、数据和策略实现，目标是为平台各参与角色提供高效易用的系统，持续迭代架构，保证稳定、安全、可扩展。在此基础上希望以技术服务并驱动业务发展，早日实现日均亿单的业务目标。广告组负责外卖餐饮和新零售流量等多元流量变现工作，是支持公司业务发展的核心保障之一。

- 岗位描述

负责强化学习和运筹优化前沿算法的创新研究与探索，发表顶会论文和申请专利
研发适用于广告场景的深度强化学习算法，如广告智能出价、门店预算分配等算法
负责深度强化学习算法的模型开发、调试

- 岗位要求

2022年毕业
计算机或相关专业硕士以上学历，保持对领域最前沿技术的追踪
能熟练使用主流深度学习框架，如tensorflow、pytorch等，具备实现常用的（深度）强化学习算法能力
在人工智能会议和期刊发表过优秀论文，有顶级会议期刊发表经历者优先（NIPS, IJCAI, AAAI, ICML, ICLR，AAMAS等）
熟悉强化学习基本算法，使用过TRPO，BCQ，REM，CQL等算法者优先

- 联系方式

微信：tangbo4909
邮件：tangbo17@meituan.com
工作地：北京

西安交通大学
XI'AN JIAOTONG UNIVERSITY